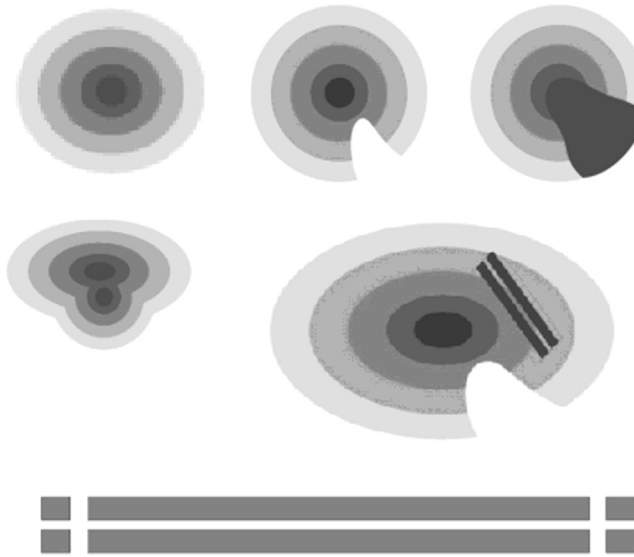
CHAPTER 3

Information encoded in protein structure

Leszek Konieczny[1], **Irena Roterman**[2]
[1]Chair of Medical Biochemistry, Jagiellonian University – Medical College, Krakow, Poland
[2]Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Krakow, Poland

Contents

References 38



Schematic depiction of variable quantities of information encoded as specific deformations in a protein "micelle" (of varying complexity), producing many different conformations – from spherical to ribbon-like. Except for the first and last structure, each form in the sequence encodes information in proportion to its deviation from

From Globular Proteins to Amyloids
ISBN: 978-0-08-102981-7
https://doi.org/10.1016/B978-0-08-102981-7.00003-8

the theoretical distribution of hydrophobicity. Such deviations may be regarded as a way to ensure the protein's specificity, (note, however that the first structure — i.e. a spherical micelle — and the last structure — a ribbonlike micelle — are devoid of information and therefore nonspecific).

The prevailing dogma which assumes that the protein's conformation is fully encoded in its sequence, seems to underestimate the role of external factors which affect protein folding.

A basic definition, due to Shannon, specifies the quantity of information (bit) carried by an event with probability p_i as:

$$I_i = -\log_2 p_i$$

This formula may be applied to information processed by biological systems. Given four types of nucleotides which encode for 20 amino acids, it is clear that — in order to unambiguously identify a specific amino acid — at least three nucleotides are required. A simple calculation then reveals that a surplus of information exists on the side of nucleotide triplets. The transfer of information between nucleotides and amino acid sequences is therefore relatively straightforward. The same, however, is not true for the transition between amino acid chains and 3D structures. The 3D structure itself may be treated as a specific means of encoding information which is crucial for the protein to perform its function. A protein's 3D structure can be accurately described in terms of dihedral angles (Φ and Ψ) between each pair of adjacent residues.

As it turns out, the quantity of information required to unambiguously define each angle with an accuracy of at least 5 degrees, is two to three times greater than the quantity carried by each amino acid. More specifically, the information content of a single residue is on the order of 4—6 bits (depending on its frequency of occurrence). In turn, the amount of information necessary to define the pair of Φ and Ψ angles with a precision of 5 degrees is 8—11.5 bits (taking into account the conformational preferences of each residue).

As shown above, there is a notable deficiency of information on the side of the amino acid chain — however, if we restrict our search to identifying secondary conformational characteristics, the available information might prove sufficient. Thus, instead of the full Ramachandran plot, we confine our search to a specific subspace, represented by an elliptical path which traverses all areas corresponding to well-known secondary folds [1,2]. This process reduces our demand for information and indicates that the input chain may indeed provide enough information to determine a conformation — but

only of the early-stage intermediate (ES). Consequently, we divide the folding process into several intermediate stages, as follows:

$$U \rightarrow ES \rightarrow LS \rightarrow \text{native 3D}$$

U – unfolded structure (plain amino acid sequence); ES – early-stage intermediate restricted to the limited conformational subspace (details in Refs. [2,3]); LS – late-stage intermediate leading to the native (3D) form of the protein.

Introducing two intermediate steps decreases the amount of information required at each stage. Restricting our search to a limited conformational sub-space, i.e. an elliptical path on Ramachandran map (see Refs. [2,3] for a detailed description), facilitates identification of Φ and Ψ angles (with a 5-degree step along the elliptical path) for the ES intermediate. It appears that the amount of information required to define these values of Φ and Ψ corresponds to the information content of the amino acid chain. This implies that the quantity of information carried by the amino acid sequence is only sufficient to determine the structure of ES. This intermediate is very important since it embodies the greatest challenge faced by protein structure prediction algorithms. Attempts to define starting structural forms (ES intermediate) in protein structure prediction models follow many diverse approaches, including the following:

1. comparative modeling: querying structure databases for 3D forms whose corresponding sequences are a good match for the given input sequence, and then applying genetic algorithms to further align the structures of both models. This method is often limited to homological proteins and can be described as evolution-based.
2. compiling databases of short structural motifs which may be assembled into a starting structure which is then subjected to further modeling. For example, the Robetta software uses lists of 3- and 9-residue fragments [4],
3. optimizing the conformational space for rapid searching by limiting the degrees of freedom available for each rotation [5],
4. simplifying the input structure by reducing it to a coarse-grained form – much like the preceding method, this operation also reduces the dimensionality of the conformational space [6].

Traversing the full conformational subspace has recently become a feasible option thanks to major advancements in IT and computer science; however this method is sometimes criticized as being out of touch with

experimental realities. Clearly, *in vivo* protein folding cannot rely on a brute-force approach, since it would then require far more time than is actually the case.

As previously noted, the information carried by the residue sequence, regardless of the applied theoretical approach, is not sufficient to unambiguously define the native 3D form of the resulting protein. A simple corollary is that progressing from ES to LS calls for an additional source of information. In our view this source is represented by the aqueous environment. It should be noted that the presence of water is an indispensable condition of proper protein folding — a fact so obvious that its full implications are often overlooked.

Attempts to factor the aqueous solvent into *in silico* folding simulations have a long history [7]. A popular approach is to introduce a certain number of water molecules (depending on solvent density) into the bounding box which contains the target chain. In this technique, the polypeptide interacts with the solvent in a pairwise fashion, i.e. through a network of atom-atom interactions (note that water may be modeled in various ways, as mono-, bi- or tri-atomic molecules). The simulation then takes into account nonbonding interactions (electrostatic, van der Waals, torsion potentials and H-bond potential) [4—6,8].

It appears, however, that the abovementioned procedure does not accurately capture the impact of the aqueous environment. Restricting analysis to a set of interactions between pairs of atoms neglects the holistic influence of the solvent upon the resulting 3D structure. Therefore, the fuzzy oil drop model disposes with this schema in favor of a continuous force field, treated as a "background" for interactions which occur between the polypeptide's constituent atoms. The specific nature of this external field depends on the structural properties of water (which, as yet, are poorly understood — at least for the liquid phase); however in all cases it promotes internalization of hydrophobic residues and formation of a hydrophobic core. A classical theory constructed on the basis of this assumption is the so-called "oil drop" model [9], which predicts the existence of a hydrophobic core encapsulated by a hydrophilic shell. In its basic version the model is discrete, i.e. it only recognizes two possible states (hydrophobic center + polar surface); however despite this drawback it accurately captures the role of the environment in terms of isolating hydrophobic residues from contact with water. We may speculate that similar results could be obtained using the pairwise interaction method; however, this would likely entail an arduous and lengthy simulation process.

In the course of our work we have proposed a continuous extension of the discrete oil drop model, which we refer to as the "fuzzy oil drop" model (Chapter 1 and 2). It enables us to quantitatively define the structural ordering of the protein's hydrophobic core, as well as to assess any potential deviations from theoretical predictions. As noted in Ref. [10], many biologically active proteins are highly consistent with the 3D Gaussian distribution of hydrophobicity, lending support for the model itself. Analysis of a nonredundant set of protein structures derived from PDB further indicates that a vast majority of individual domains adhere to the model with high accuracy [10].

Many protein families, including the so-called downhill and fast-folding proteins [11,12], fold in accordance with the fuzzy oil drop model. This shows that the quantity of information present in the amino acid sequence is sufficient to determine its 3D structure — but only in the presence of an additional source of information, i.e. the solvent. The fast-folding family is particularly noteworthy in this respect: its members are capable of rapidly reverting to their native forms regardless of how many times they have been unfolded. This ability to "automatically" assume the intended tertiary conformation in the absence of any other stimuli clearly shows that missing information comes from the solvent (even without assessing the specific quantity of bits which the solvent imparts to the protein).

For the reasons stated above we feel confident in stating that the ES-to-LS folding stage draws information from the protein's environment, and that therefore the environment plays an active and crucial role in the folding process. Interestingly, experimental studies report that undesirable changes in environmental properties (changes in pH, ionic potential etc.) have a detrimental effect on protein folding — an observation which is fully consistent with our proposed theory.

The equivalence between the quantity of information fed into the folding process (residue sequence; aqueous environment) and the information content of the resulting 3D structure is not so much computed as experimentally demonstrated by reversibility of folding/unfolding. Still, this equivalence does not fully resolve the question of how proteins attain their intended conformations. The presented approach describes a general process without referring to the function of specific proteins. In some cases no function-oriented structural changes are necessary — for example, antifreeze proteins (which are structurally highly consistent with the fuzzy oil drop model) perform their function simply by being dispersed in the solvent where they can disrupt the formation of ice crystals. Such proteins do not

need to attract or bind to any external structures since this would restrict their exposure to water.

In most cases, however, the schematic folding process presented above omits a crucial issue: the way in which the protein's 3D structure encodes its intended function:

$$U \rightarrow ES \rightarrow LS \rightarrow 3D \rightarrow FUNCTION$$

This additional stage – from 3D structure to function – may also be analyzed on the basis of information theory concepts.

A hypothetical protein which is fully consistent with the 3D Gaussian distribution of hydrophobicity would exhibit two important properties: perfect solubility and inability to interact with any other molecules (except for surface-bound ions). In actual proteins whose activity calls for interaction with external structures, certain local deviations from the theoretical distribution of hydrophobicity are expected. As shown in the chapter devoted to ligand binding and protein complexation (Chapter 6), the nature of this phenomenon is highly complex. Most proteins are very selective and therefore need to deviate from the theoretical distribution of hydrophobicity in a very specific and controlled manner. Thus, the quantity of information needed to produce a protein with a specified activity profile must be greater than the quantity needed to produce an inert protein (or a protein whose activity is entirely determined by its solubility). This, again, calls for an additional source of information.

In our previous book which introduces the fuzzy oil drop model, we postulated [13] that the enzyme must fold in the presence of its target substrate. The substrate would need to take an active part in the folding process – the enzyme effectively folds "around" the substrate, which automatically generates a suitable binding cavity. Note, however, that this assumption remains speculative and calls for experimental validation.

There are, however, other potential sources of information, such as chaperones: proteins, which work by temporarily attaching themselves to the target polypeptide chain and modulating the external force field which guides the folding process. It is also worth noting that any such modulations are local in scope and produce similarly local distortions in the resulting structures.

The quantity of information carried by a chaperone may be precisely calculated by comparing the structure of the chaperone-assisted protein with the hypothetical conformation which it would reach in the absence of the chaperone. Kullback-Leibler's divergence entropy coefficient, D_{KL}, provides a *de facto* quantitative measure of this difference [14].

Can a protein be accurately referred to as a chemical molecule? Or, in other words — what differentiates a protein from any other compound, whether organic or inorganic? Saying that "the protein is synthesized by a living organism" does not fully address the problem. Instead, we propose a definition which refers to the targeted nature of proteins. The protein's native form is not a goal unto itself, and the folding process does not end when the protein attains its intended 3D structure — rather, the process may be considered complete only after the protein has gained biological function. In this sense, the term "folding" relates not to the protein's structure, but to its biological role. If the protein has not folded in the intended manner, it is degraded and disposed of: from the biological point of view it is useless and cannot be considered a "biological molecule", even though it may have been synthesized by a living organism.

In short, the protein is a tool which must perform a specific task. It would therefore be misleading to assume that the 3D structure is the ultimate goal of the folding process.

The sources of information required to produce specific local deviations from the ideal distribution of hydrophobicity are varied and depend on the complexity of the structure which must be produced to perform the given task. Validating the correctness of simulated structures makes sense only if the protein in question has been proven to perform its function. Consequently, RMS-D scores should not be regarded as the sole criterion of the reliability of structure prediction algorithms [15,16].

Many evolutionarily conditioned processes which produce desirable results may be explained by invoking the following formula, well known to information scientists:

$$P = [1 - (1 - p)k]$$

P — overall probability of successfully completing a task; p — probability of success for a single attempt; k — number of attempts.

There are two ways to increase P (i.e. the odds of successfully completing a given task). One way to approach this problem is to maximize k, i.e. the number of attempts, each of which may succeed with probability p. A classic example is a lottery where each contestant may purchase an arbitrary number of tickets. Purchasing more tickets (greater k) gives one better chances of winning the lottery (greater P). This method is most often applied when the player cannot deduce the correct solution *a prori* (i.e. when the winning numbers are not known to them). The downside is that it entails a significant expenditure of resource and energy to produce the large number of "coins" whatever its form is.

The other approach is to try to increase the value of p. In the case of a numbers game we may achieve this effect by investigating how the winning numbers are drawn, and by introducing a deterministic factor (for example, if the drawing machine uses numbered balls, we may surreptitiously insert small magnets into certain balls and a larger magnet into the drawing tube, thus increasing the likelihood that our designated balls will be selected). In this case, the expenditure of resources is negligible, but the process instead calls for additional information — specifically, information regarding the inner workings of the selection process. In short, we must be aware of how the system works and also know the desired outcome (set of balls), both of which require information.

In biology a classic mechanism which relies on high values of k is plant pollination. Since a plant is unable to determine the optimal placement of seeds to ensure germination, plants instead produce vast quantities of seeds and must bear the associated costs (energy expenditures). Similar solutions are employed by the reproductive systems of animals, including mammals. A sperm cell does not know how to locate the egg — it therefore faces a similar problem to a plant trying to disseminate its offspring.

In contrast, some plants produce rhizomes, which represent an alternative way of increasing the likelihood of successful proliferation — in this case, by increasing the value of p. The rhizome is essentially a living laboratory, which takes care of all of the plant's vital processes. A robust rhizome serves as proof that the plant's requirements are well taken care of. Such rhizomes are more likely to sprout a new plant. When digging out a plot of perennials we may sometimes find wilted, stunted rhizomes which have not encountered suitable conditions and cannot initiate proliferation — or even support themselves. Thus, the condition of the rhizome provides additional information which is required to increase the value of p. In this particular case the required quantity of information is vast and involves all vital processes which take place in the rhizome. Notably, the rhizome also encodes the intended outcome of the selection process: finding a place which promotes growth of its offspring. A human analogy would be *in vitro* fertilization — thanks to in-depth knowledge of the reproduction process and also of its expected outcome, we may create a new organism using just a handful of sperm cells.

On the molecular level, the "increased k" approach is embodied by synthesis of IgG antibodies. This process proceeds without knowledge of the intended target, i.e. the antigen which may have entered the organism. Modifications of CDR fragments increase the likelihood of obtaining a

combination which matches a particular (but unknown) antigen. The greater the diversity of IgG antibodies the greater the chance of initiating a successful immune response. Of course, the flipside of this process is that it consumes a great deal of energy, needed to sustain synthesis of a vast array of proteins which differ with respect to their CDR fragments. In contrast, vaccination represents a way to tackle the problem by increasing the value of p, since the antigen (representing the given disease) has already been recognized and may be introduced into the organism in a controlled manner, to develop immunity.

Approaches based on increasing p are also employed in systems which encode information related to the specific nature of their associated processes, such as all enzymatic catalysis reactions.

The quantity of information encoded in any specific system varies along with the system's complexity and its subdivision into stages. For example, enzymatic active sites, responsible for catalysis and admitting the presence of water, are usually found on the molecular surface or in shallow pockets. If, however, a given reaction requires an anhydrous environment, the system becomes far more complex since the active site must accommodate its intended substrate while at the same time excluding water. Examples of highly complex systems which rely on the "increased p" approach are provided by all enzymes (or receptors) which have a quaternary structure. This layer of the structural hierarchy resolves two problems: it facilitates construction of an active site characterized by high information content (complicated structure), and it also provides a way to prevent unintended initiation of a given process. The latter property is ensured by the complex nature of the receptor itself: if any one of its components is missing, the process cannot begin. In some cases, the number of required components (and therefore the quantity of required information) is very high. An extreme example is provided by the ribosome, which, on the one hand, must know how to carry out protein synthesis, while on the other hand must be able to validate that all conditions for the synthesis of a given protein have been met (i.e. that all required components are present). This complicated process also provides a way to exert tight control over protein synthesis, which is fundamentally important for all organisms.

The visualization of the P dependence on p and/or k is shown in Fig. 3.1.

The complex form of a ribosome (as wall as of any receptor which has a quaternary structure) encodes a lot of information. In theory, it might be possible to pack the necessary information into a single polypeptide chain. In practice, however, synthesizing a complex active site becomes far easier

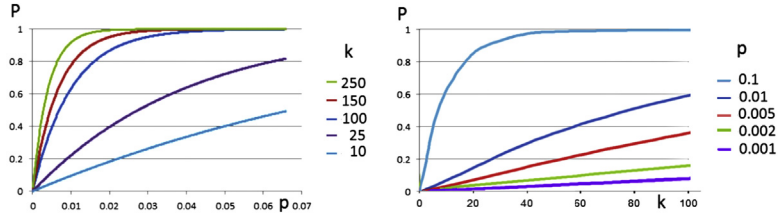


Fig. 3.1 Association between the probability of reaching the goal (*P*) and: (Left) the number of attempts (*k*) for a variable elementary probability (*p*), (Right) the probability associated with each elementary event (*p*) for a variable number of attempts (*k*).

when the given structure is assembled from smaller fragments, each of which contributes some of the necessary information. This assembly provides a way to combine small pieces into a single information-rich system whose operation may be compared (on information theory grounds) to a logical conjunction of events. Notably, conjunction carries more information than a logical alternative (which would be difficult to implement within the bounds of a single biological structure) — this is why the active site is often located in closeness, proximity to the molecular (or domain) interface, or even forms part of that interface. Each structural unit contributes a piece of information required to trigger a complex and highly specific process.

Enzymatic catalysis, in addition to providing the necessary conditions for a given reaction to occur, must overcome one additional obstacle: the need for intermolecular communication. The protein's structure (including its quaternary structure) must, in addition to enabling catalysis, also send out specific signals to attract potential partners, i.e. other molecules involved in the given reaction. Such short-range intermolecular communication needs to make use of the aqueous environment. This issue will be further discussed in Chapter 7, where we investigate the effects exerted by the protein upon the surrounding solvent.

Let us refer once again to the definition of a protein as a tool required to perform a specific and potentially highly complex task. A chemical molecule obtained through artificial synthesis (whether organic or inorganic) lacks an inbuilt purpose — it simply exists. The same is true for a protein devoid of biological activity: it may be regarded as a standalone entity, not associated with any process and not fulfilling any goal. Such anomalous proteins, often resulting from mutations, are undesirable and, in most cases, disposed of by the organism. A particularly interesting example of an aberrant protein is supplied by amyloids (naturally, we refer to pathogenic amyloids, rather

than to amyloid-like molecules which the organism can make use of — the differences between both classes are presented in chapter 7).

A properly folded protein, capable of performing its intended task (whatever it may be), encodes information related to that task. In some cases, the protein's function implies interaction with another molecule, membrane or cell — the nature of the interaction partner must therefore also be encoded.

Can a micelle function as a carrier of information? A surfactant micelle, much like a micellar protein (e.g. a type III antifreeze protein), emerges as a result of immersion in an aqueous environment, which determines its structure through selective interaction with polar groups. The micelle therefore exemplifies a response to an external force field, generated by water. It encodes as much information as is necessary to progress from ES to LS. This information is encoded in the structure of the monomer (surfactant) — particularly in its size and polarity gradient, both of which determine its reaction to the external force field.

The structure of a micelle may be characterized as a passive adaptation to external forces, explaining the similar solubility of surfactant micelles and type III antifreeze proteins [17,18].

How, then, should we define an amyloid? Amyloids are essentially anomalous ribbon-like or cylindrical [19] micelles, which emerge not as a *result* of external forces, but *in spite* of them. The aqueous solvent does not control the process which leads to formation of amyloid fibrils — rather, the fibril actively opposes the influence of the solvent. The source of information may be found in the intrinsic properties of constituent residues rather than in their environment. This observation may be regarded as paradoxical, but it explains the peculiar properties of amyloids. In their case, information is carried by the sequence — or, more accurately, by the intrinsic hydrophobicity of its constituent residues. It appears that amyloid structures emerge when the effect of the external force field is diminished, for instance by a change in the physiochemical properties of water. This corresponds to a change in the properties of the external field, enabling intrinsic hydrophobicity to guide the folding process to a conclusion which differs from "natural" conditions. Notably, physiochemical changes in the solvent have been found to promote amyloidogenesis [19].

Shaking is a known "nonchemical" inducer of amyloid formation. In physical terms, shaking alters the properties of the solvent through aeration, which, in turn, increases the interphase boundary area. The folding process is also hindered in the presence of detergents — likely not through direct

interactions with the polypeptide chain but rather due to a change in the properties of water [19].

To conclude our study of the role of information in protein folding/misfolding, we may remark that the aqueous environment — when modeled as a continuous force field — provides information crucial for proper folding of polypeptide chains and ensuring that they are capable of fulfilling their biological role. Changes limited to the properties of this field (e.g. presence of urea) may result in protein disability. Very often the process is reversible and in the absence of the denaturing factor the solvent reverts to its natural state, where proper folding may again occur. Unfortunately, however, amyloids do not follow the above rules: once formed, they do not undergo structural modifications when environmental conditions change.

The above hypothesis constitutes a core aspect of the presented study.

References

[1] Alejster P, Jurkowski W, Roterman-Konieczna I. Structural information involved in the interpretation of the stepwise protein folding process. Protein Folding in Silico 2012:39—54. https://doi.org/10.1533/9781908818256.39.

[2] Jurkowski W, Brylinski M, Konieczny L, Wiíniowski Z, Roterman I. Conformational subspace in simulation of early-stage protein folding. Proteins: Structure, Function, and Bioinformatics 2004;55(1):115—27. https://doi.org/10.1002/prot.20002.

[3] Jurkowski W, Baster Z, Dułak D, Roterman-Konieczna I. The early-stage intermediate. Protein Folding in Silico 2012:1—20. https://doi.org/10.1533/9781908818256.1.

[4] Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research 2004;32(Web Server):W526—31. https://doi.org/10.1093/nar/gkh468.

[5] Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins: Structure, Function, and Genetics 1994;18(4):338—52. https://doi.org/10.1002/prot.340180405.

[6] Krupa P, Hałabis A, Żmudzińska W, Ołdziej S, Scheraga HA, Liwo A. Maximum likelihood calibration of the UNRES force field for simulation of protein structure and dynamics. Journal of Chemical Information and Modeling 2017;57(9):2364—77. https://doi.org/10.1021/acs.jcim.7b00254.

[7] Science (n.d.). doi:10.1126/science.

[8] Van Gunsteren WF, Berendsen HJC. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. Angewandte Chemie International Edition in English 1990;29(9):992—1023. https://doi.org/10.1002/anie.199009921.

[9] Kauzmann W. Some factors in the interpretation of protein denaturation. Advances in Protein Chemistry 1959;14:1—63. https://doi.org/10.1016/s0065-3233(08)60608-7.

[10] Sałapa K, Kalinowska B, Jadczyk T, Roterman I. Measurement of hydrophobicity distribution in proteins - non-redundant protein data bank. Bio-Algorithms and Med-Systems 2012;8. https://doi.org/10.2478/bams-2012-0023.

[11] Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M. Two-intermediate model to characterize the structure of fast-folding proteins. Journal of Theoretical Biology 2011;283(1):60—70. https://doi.org/10.1016/j.jtbi.2011.05.027.

[12] Kalinowska B, Banach M, Wiśniowski Z, Konieczny L, Roterman I. Is the hydrophobic core a universal structural element in proteins? Journal of Molecular Modeling 2017;23(7). https://doi.org/10.1007/s00894-017-3367-z.
[13] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics 1951;22(1):79–86. https://doi.org/10.1214/aoms/1177729694.
[14] http://predictioncenter.org.
[15] Gadzała M, Kalinowska B, Banach M, Konieczny L, Roterman I. Determining protein similarity by comparing hydrophobic core structure. Heliyon 2017;3(2):e00235. https://doi.org/10.1016/j.heliyon.2017.e00235.
[16] Roterman I, Banach M, Konieczny L. Antifreeze proteins. Bioinformation 2017; 13(12):400–1. https://doi.org/10.6026/97320630013400.
[17] Banach M, Konieczny L, Roterman I. Why do antifreeze proteins require a solenoid? Biochimie 2018;144:74–84. https://doi.org/10.1016/j.biochi.2017.10.011.
[18] Roterman I, Banach M, Konieczny L. Application of the fuzzy oil drop model describes amyloid as a ribbonlike micelle. Entropy 2017;19(4):167. https://doi.org/10.3390/e19040167.
[19] Serpell LC. Alzheimer's amyloid fibrils: structure and assembly. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 2000;1502(1):16–30. https://doi.org/10.1016/s0925-4439(00)00029-6.